Nature de l'information exploitée pour identifier perceptivement une voyelle en parole continue

Michel Pitermann

Laboratoire Parole et Langage, URA 261 CNRS
Université de Provence, 29 av. Robert Schuman
13621 Aix-en-Provence Cedex, France
mpiter@lpl.univ-aix.fr

SUMMARY

This paper presents a perceptual experiment on stimuli synthesized by means of a vocal tract model. The purpose was to find out the nature of the main piece of information used to identify a vowel in continuous speech. The information contained in three parts of the speech signal was compared : (i) the vowel nucleus ; (ii) the acoustical contrast between the vowel nucleus and its context ; (iii) the transitions between the "stationary" parts of the speech signal. Results show that the vocoïds were better identified by means of dynamic than static or acoustical contrast information. However, to be generalized, this conclusion must be confirmed by other experiments.

1. INTRODUCTION

La nature de l'information que notre système perceptif traite pour identifier une voyelle en parole continue reste controversée. Parmi les différents types d'information généralement étudiés, citons trois exemples : (i) l'information intrinsèque à la voyelle contenue dans son noyau vocalique (Assmann et al. 1982) ; (ii) le contraste acoustique existant entre la partie quasi stationnaire de la voyelle et celle de son contexte (Nearey 1989) ; (iii) l'information dynamique contenue dans les transitions reliant les parties stables du signal de parole (Carré et al. 1994).

Quelques expériences de perception à partir de stimuli de parole synthétisés à l'aide d'un modèle du conduit vocal ont été présentées dans (Carré et al. 1994). L'une des expériences avait pour objectif de comparer le rôle de la partie quasi stationnaire de la voyelle à celui des transitions comprises entre les segments phonétiques pour la catégoriser perceptivement. Les vocoïdes synthétiques contenant de l'information dynamique ont été mieux catégorisés, mais le contraste acoustique entre segments voisins pouvait expliquer ce résultat. Nous avons donc étendu l'expérience afin de déterminer la contribution du contraste acoustique dans ce type d'expérience.

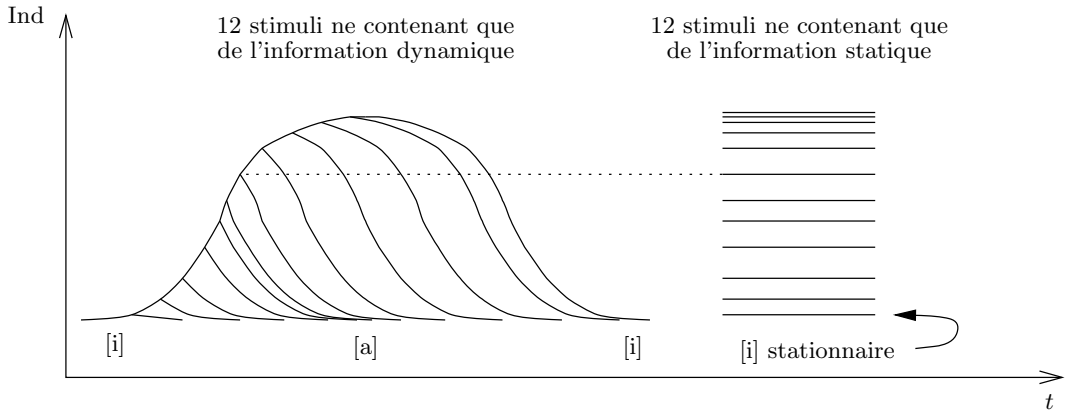

FIG. 1: Construction de vocoïdes. Le graphique représente un indice fictif prenant une valeur basse ou haute respectivement pour la production d'un [i] ou d'un [a]. Pour forger un vocoïde stationnaire, un vecteur de paramètres de commandes du modèle prélevé dans la transition [ia] a été dupliqué un nombre suffisant de fois pour synthétiser un signal stationnaire de 300 ms (cf. partie droite de la Figure). Le [i] stationnaire de 300 ms utilisé comme contexte a été conçu grâce au premier vecteur de la transition [ia]. Pour construire un stimulus [iVi] ne contenant pas d'information statique, une coordonnée temporelle de la transition [ia] était sélectionnée. Le segment de transition [iV] la précédant était dupliqué en miroir et ajouté au segment de transition initial afin de créer le stimulus [iVi] (cf. partie gauche de la Figure).

2. MÉTHODE

Grâce à un modèle du conduit vocal, des vocoïdes ne contenant que de l'information statique ou dynamique ont été synthétisés. Des auditeurs inexpérimentés ont dû les identifier perceptivement. Le déplacement des frontières de catégorisation en fonction de la nature de l'information utilisée pour la synthèse renseigne sur les contributions relatives de ces types d'information pour la perception des vocoïdes.

Le modèle du conduit vocal était celui de Kelly-Lochbaum à six tubes (Schoentgen and Ciocea 1995). Cinq des six sections et la longueur totale constituaient les paramètres de commandes.

Un signal de parole naturelle [iai] a servi de point de départ pour la construction des stimuli. Les commandes correspondantes du modèle de conduit vocal ont été estimées par inversion acoustico-articulatoire (Schoentgen and Ciocea 1995). La Figure 1 illustre comment de nouvelles commandes du modèle ont été générées pour produire trois classes de stimuli : (i) des vocoïdes stationnaires isolés (stimuli [V]) ; (ii) des vocoïdes stationnaires placés entre deux [i] soutenus générant un contraste acoustique (stimuli [i#V#i], où # désigne un silence de 20 ms) ; (iii) des vocoïdes construits uniquement à partir des transitions intervocaliques (stimuli [iVi]).

Pour synthétiser un signal acoustique à partir d'une commande du modèle du conduit vocal, les séries chronologiques correspondantes des trois premiers formants étaient calculées (Schoentgen and Ciocea 1995). Le synthétiseur à formants de Klatt a alors été utilisé pour produire le signal acoustique (Klatt and Klatt 1990).

La transition [ia] originale contenait 12 vecteurs de paramètres de commandes du

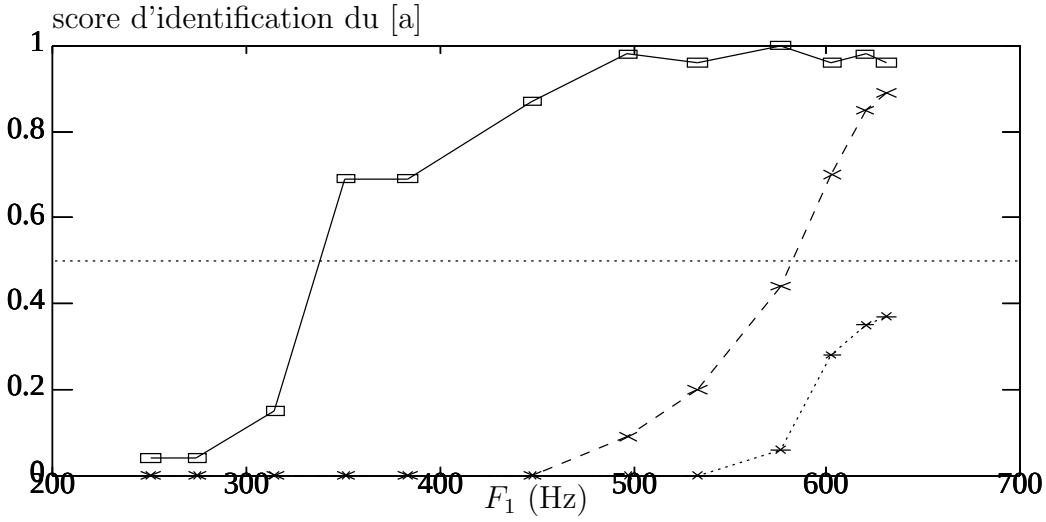

FIG. 2: Proportion de stimuli catégorisés [a] en fonction de la valeur du premier formant. Les signes '□', '×' et '∗' correspondent respectivement aux stimuli [iVi], [i#V#i] et [V]. La ligne horizontale pointillée indique la proportion 0.5 utilisée pour estimer les frontières de catégorisation entre le [a] et le [ə].

modèle. Par conséquent, trois ensembles de 12 stimuli ont été produits. Chaque signal acoustique a été dupliqué 6 fois afin de créer une liste de 72 vocoïdes stationnaires isolés [V] et une liste de 144 stimuli [i#V#i] et [iVi] mélangés.

Neuf auditeurs francophones inexpérimentés âgés de 22 à 30 ans ont écouté les deux listes de stimuli. Les signaux ont été présentés dans un ordre aléatoire différent pour chaque sujet afin de limiter les erreurs d'identification liées à l'ordre de présentation. Cinq auditeurs ont commencé par les stimuli [V], les quatre autres par l'autre liste. Un silence de trois secondes séparait les signaux successifs.

Les instructions étaient de retranscrire sur une feuille de papier la voyelle soutenue isolée ou la voyelle comprise entre les deux [i]. Aucune indication sur le type de voyelle à reconnaître ou sur la transcription phonétique à utiliser n'a été donnée.

3. RÉSULTATS

Sur le continuum [a]–[i], les sujets ont catégorisé les vocoïdes dans l'ordre [a], [ə], [ɛ], [e] et [i]. La présence des [ə] s'explique par le fait que la transition [ia] passait par la valeur du triplet des trois premiers formants (500,1650,2470) (valeurs en Hz).

La Figure 2 montre la proportion de stimuli catégorisés [a] en fonction de F_1.

4. DISCUSSION

Les résultats sont homogènes par rapport aux auditeurs. La seule exception est le groupe des trois vocoïdes stationnaires isolés correspondant aux valeurs du premier

formant les plus élevées (les 3 signes '$*$' de droite la Figure 2). Pour ces trois vocoïdes, les sujets ont été divisés en deux groupes. Cinq auditeurs ont systématiquement catégorisé ces vocoïdes [ə], les autres ont plutôt perçu [a].

Les résultats indiquent que les vocoïdes stationnaires isolés ont été peu catégorisés [a]. Par conséquent, le noyau vocalique du [a] de départ ne semble pas correspondre à un [a] de bonne qualité.

La Figure 2 montre que le contexte [i#_#i] modifie la perception des vocoïdes synthétiques, mais ne suffit pas à transformer leur qualité en celle d'un bon [a].

Par contre, les résultats correspondant aux vocoïdes ne contenant que de l'information dynamique se distinguent nettement. Sept vocoïdes parmi les douze ont été clairement perçus [a]. Même lorsque le triplet (F_1,F_2,F_3) n'atteignait que (450,1700, 2500) (valeurs en Hz), 87 % des vocoïdes ont été catégorisés [a]. Le déplacement de la frontière de catégorisation des stimuli [iVi] par rapport aux stimuli [i#V#i] est de -240 Hz pour F_1 et 250 Hz pour F_2.

5. CONCLUSION

L'identité perçue des vocoïdes synthétisés pour cette expérience dépendait essentiellement de la dynamique des déformations du modèle du conduit vocal. Pour être généralisé, ce résultat doit être confirmé par d'autres expériences.

6. REMERCIEMENTS

Ce travail a été réalisé à l'Institut des Langues Vivantes et de Phonétique de Bruxelles.

BIBLIOGRAPHIE

Assmann, P., T. Nearey, and J. Hogan (1982). Vowel identification : Orthographic, perceptual and acoustic aspects. *The Journal of the Acoustical Society of America 71*, 975–989.

Carré, R., S. Chennoukh, B. Lindblom, and P. Divenyi (1994). On the perceptual caracteristics of "speech gestures". *The Journal of the Acoustical Society of America 96*, 3326–3326. ASA meeting held in Austin.

Klatt, D. and L. Klatt (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America 87*(2), 820–857.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America 85*(5), 2088–2113.

Schoentgen, J. and S. Ciocea (1995). Direct calculation of the vocal tract area function from measured formant frequencies. In *Eurospeech'95 Proceedings*, Volume 1, Madrid, Spain, pp. 745–748. European Speech Communication Association.